GeLexi projekt: GEneratív LEXIkonon alapuló mondatelemzés

Alberti Gábor, Kleiber Judit, Viszket Anita

Pécsi Tudományegyetem, BTK, Nyelvtudományi Tanszék
H7624 Pécs, Ifjúság útja 6., Magyarország
gelexi@btk.pte.hu
http://lingua.btk.pte.hu/gelexi.asp

Kivonat. A 2001-ben Pécsett alakult *GeLexi* kutatócsoportunk[1] kiinduló célja a számítógépes implementálás révén való legitimálása egy olyan grammatikának, amely a generatív nyelvészeti kutatásokban a hatvanas évektől egyre erősödő *lexikalista* tendenciát a „végsőkig" fokozza: lemond a frázisstruktúra-építésről, miközben a szórendről mégis számot ad, rangsorolt szomszédossági követelmények segítségével. Másik újdonsága a „totális lexikalizmus" kiterjesztése a morfológiára is: nem a szavakhoz, hanem a morfémákhoz tartoznak a minden grammatikai szintről egyidejűleg információt hordozó lexikai egységek, amelyek oly módon „szerelik össze magukat" a mondatelemzés során, hogy az is eldől, mely elemek épülnek össze szóvá, és melyek alkotnak (egymás „közelében" maradó) külön szavakat. Nyelvelméleti törekvéseink gyakorlati hozadéka egy folyamatos fejlesztés alatt álló program, amely egy (jelenleg magyar vagy angol) szósoron elvégzi a „generatív grammatikai alapfeladatokat": eldönti, hogy jól formált szavakból álló grammatikus mondattal áll-e szemben, majd (a pozitív esetben) kétféle diskurzus-szemantikai elemzést kínál. E cikkben *és* kötőszavas mondatok elemzése szemlélteti majd eljárásainkat.

1 Bevezetés

Általános célunk annak igazolása, hogy a költség–haszon elvet mindenek fölébe helyező, „sekélyelemző" szemléleten érdemes lenne túllépnie a számítógépes nyelvészetnek, visszafordulva a tiszta nyelvelméleti alapok felé. Kidolgozható ugyanis olyan formális (generatív) grammatika [4], amely éppen a modern számítástechnikában előnyösnek mondott kapacitásmegosztást mutatja: „minimális processzálás – maximális adattár".[2]

[1] A cikk megírását és a szegedi konferencián való jelenlétünket a T 38386 számú OTKA pályázat, valamint az első szerző esetében a MTA Nyelvtudományi Intézetének Hajdú Péter Vendégkutatói Ösztöndíja tette lehetővé. Köszönettel tartozunk továbbá Balogh Katának, aki a jelenlegi mondatelemző szoftverünk szintaktikai és szemantikai részeinek a nagy részét írta, de sajnos már nem tagja a GeLexinek, mivel (szerencsére) amszterdami doktoranduszhallgatóként intenzív szemantikaelméleti kutatásokba fogott.

[2] Összahangban azzal a lehetőséggel, amelyről Prószéky (persze a MorphoLogic gépi fordítási projektuma kapcsán) így ír: „a memóriakapacitás korábban nem tette lehetővé ilyen számú és méretű minta egyidejű használatát". Közben a kezdetekben [15] maximálisan „processzálás-párti"

Kutatócsoportunk kiinduló célja (ld. a 2. pontot) az előbbiekben közölt céllal nagy átfedést mutat, lényegében annak az elméleti nyelvész szemszögéből való megfogalmazásáról van szó: a számítógépes implementálás révén kívántunk legitimálni egy olyan (homogén felépítéséből adódóan metaelméleti érdekességgel is bíró [4]) grammatikát (GASG: *Generatív argumentumszerkezet-grammatika*), amely az iménti lábjegyzetben említett *lexikalista* tendenciát a „végsőkig" fokozza: lemond a frázisstruktúra-építésről, miközben a szórendről mégis számot ad, rangsorolt szomszédossági követelmények segítségével (egyesítve az eddigiekben meghivatkozott frázisstruktúra-nyelvtanok előnyeit a függőségi nyelvtanokéival [23]).

A 3. pontban a kutatócsoport hároméves tevékenységét tekintjük át, azzal a feltételezéssel élve, hogy ez az első magyar számítógépes nyelvészeti konferencia elsősorban a műhelyek bemutat(koz)ására szolgál.[3]

A 4. pontban felvillantjuk mondatelemző Prolog-programunk jelenlegi tudását, két futtatás végeredményéből idézve, különös tekintettel az *és* kötőszóra (annak a stratégiánknak a jegyében, amelynek értelmében minden koferencián bemutatunk valamilyen új eredményt túl az általános koncepció összefoglalásán).

2 A kiinduló cél

A *generatív* nyelvészetnek [15] elévülhetetlen érdeme, hogy *formális elméletet* nyújtott annak a régi felismerésnek a megragadására, miszerint a mondatjelentésnek két forrása van: a lexikai elemek és az azokat összefűző szerkezet. Kezdetben e szerkezetek kombinációs lehetőségeinek a matematikai vizsgálata jelentette azt a központi kérdést —ld. a Chomsky-féle nyelv(tan)osztályok témakörét, különös tekintettel a környezetfüggetlen és környezetfüggő grammatikákra [15, 21]—, amelyből két „gyakorlatiasabb" tudományág is sarjadt: a generatív nyelvleírásé és a számítógépes nyelvészeté. A generatív nyelvleírásban a mondatösszetevők mozgatását megengedő Chomsky-féle irányzatnak nem sikerült elméletileg igazolnia e mozgató *transzformációk* elengedhetetlenségét [21], ezért a hetvenes évektől kezdve több olyan „eretnek" generatív irányzat is meg tudott szilárdulni [13, 14, 17, 18], amely „alig" lépte túl a környezetfüggő eszköztár kereteit (Partee-ék [21] kiválóan bemutatják ezeknek az *enyhén környezetfüggő* eszköztáraknak a variációit). Ezzel összefüggésben az új irányzatok a lexikon+szintaxis egységben a lexikonnak a korábbiaknál jóval nagyobb szerepet biztosítottak a nyelvi jellegzetességek megragadásában, a szintaxisnak általában csupán nagyon általános frázisstruktúra-építő szabályokat meghagyva. A lexikalista tendencia elsőprő erejét mutatja, hogy a kilencvenes években a Chomsky-féle „fővonal" is ugyanilyen fordulatot vett [16] a Minimalista program keretében.

A bevezetésben említett GASG a lexikalizmus totálissá fokozásának kísérletéből született, megvalósítva Karttunen „radikális lexikalizmusát" [19]: olyan nyelvtan, amelyben

(azaz szintaxis-központú) generatív nyelvészet is erőteljes *lexikalista* fordulatot vett [13, 14, 16, 18, 19]. Hadd idézzük a faépítő nyelvtanok [21] „atyjának", Joshinak (2003) két aktuális mottóját [18]: „Complicate Locally, Simplify Globally", és „Grammar ≈ Lexicon".

[3] A bemutatkozó pont beiktatása azzal a káros következménnyel járt, hogy „felborult" a hivatkozásjegyzék a sajátpublikációk javára, ami amúgy szándékaink ellen való.

nemhogy transzformációs szabályok nincsenek, de még összetevős szerkezeti fák sem épülnek, ugyanakkor a generatív nyelvészeti *alapfeladat* elvégeztetik, azaz meghatározható, levezethető a jól formált mondatok pontos halmaza, a levezetés során pedig szintaktikai és szemantikai szerkezet rendelődik az illető (jól formált) mondathoz. Egy totálisan lexikalista nyelvtanban a grammatikai „tudás" a lexikai tételek leírásába — és csakis oda — van beépítve; minden egyes szó megmondja, hogy milyen „környezeti követelményeket" kell kielégítenie egy őt tartalmazó jól formált mondatnak. Az egyes lexikai tételek természetesen nemcsak a környezeti követelmények leírását tartalmazzák, hanem a „sajátszó" jellemzését is, hiszen más szavak éppen az illető szót fogják keresni potenciális mondatokban.

A (generatív) nyelvészeti leírásokat — elvi okok miatt is! — igen fontos számítógépes implementációval igazolni, hiszen a működő algoritmusok teszik kétségtelenné, hogy jól formalizált, egzakt rendszerünk van. Az implementáló programunk tevékenysége nem más, mint az imént említett „generatív alapfeladat" végrehajtása: egyértelműen el kell döntenie egy beírt magyar szósorról, hogy az grammatikus-e, vagy sem, és amennyiben az, morfoszintaktikai és szemantikai reprezentációt kell hozzá rendelnie.

Nyelvtanunk tulajdonságairól a következő pontokban még lesz szó; most a bevezetés kezdőgondolatához szeretnénk visszakanyarodni azzal a megjegyzéssel, hogy a memóriakapacitás növekedése és a mintaillesztési eljárások fejlődése a technológia oldalán [22] szerencsésen összetalálkozhat majd a lexikalizmusnak köszönhető homogén generatív nyelvelmélettel, amit katalizálhat az intelligens alkalmazások iránt felébredő —immár nem irreális— igény. A GASG gyakorlatilag legitimálhatja a nyelvtechnológiában eddig ösztönösen alkalmazott, szókörnyezetek illesztésén alapuló heurisztikus eljárásokat, elméletileg is korrekt rendszerré továbbfejlesztve azokat.

3 A GeLexi tevékenysége

A GASG számítógépes implementációjának javaslata először 1998-ban kapott nagyobb nyilvánosságot egy debreceni nemzetközi konferencián [1], bár az első szerző az „Alkalmazott Logikai Laboratórium" (ALL) munkatársaként már a kilencvenes évek elején készített néhány belső anyagot a témakörben. A korábbi elképzelések alapján 2001-ben végre elkészült egy elemző program, amely magyar szavakból álló sorozatokról eldöntötte (pusztán a toldalékolt szavak lexikai leírására támaszkodva), hogy jól formált mondatot alkotnak-e (az adott szórend és toldalékolás mellett) [9,3].

Majd megkezdtük a GASG szemantikai reprezentációjának kidolgozását. Más nem is jöhetett szóba, mint a Kamp-féle „DRS-ek" (diskurzusreprezentációs struktúrák) [17] egy továbbfejlesztett változatának [2] implementálása, mivel tulajdonképpen a GASG létjogosultsága melletti egyik fő elméleti érvet éppen abban látjuk, hogy *kompozicionális* [21] szemantikai partnereként szolgálhat az említett ígéretes (a Montague-féle szemantikai rendszereken [21] relevánsan túllépő) diskurzus-szemantikai elméletnek. Addigi elképzeléseink alapos (főleg elméleti) bemutatásával szolgál egy elvekről, modellekről és szabályokról szóló szegedi konferencia kötete [4], amely a nyelvleírási szintek közül a morfológiát nem tárgyalja alaposabban. A megelőző időszakban ugyanis a (toldalékolt szavakhoz rendelt) lexikai egységeket egy óriási *öröklődési hálózatban* gondoltuk elrendezendőnek, elkészítésüket pedig a bevett, reguláris kapacitású eszközökre [20] kívántuk bízni.

Később azonban egyértelművé vált, hogy a GASG kívánatos teljes homogenitását az biztosítja, ha a morfológiát is „totálisan lexikalista" módon közelítjük meg: nem a szavakhoz, hanem közvetlenül a morfémákhoz rendeljük a gazdagon strukturált, minden grammatikai szintről egyidejűleg információt hordozó lexikai egységeket, amelyek oly módon „szerelik össze magukat" a mondatelemzés során, hogy az is eldől, mely elemek épülnek össze szóvá, és melyek alkotnak (egymás „közelében" maradó) külön szavakat [5, 10, 11]. Ezen a ponton teszünk említést arról a fontos technikáról, amely a GASG-ben kiváltja a —szórend meghatározásáért felelős!— frázisstruktúra-építést: olyan *rangparaméteres* szomszédossági követelményeket támasztanak a mondatba kerülő lexikai egységek egymás iránt, amelyeket *közvetett módon* is ki lehet elégíteni, magasabb rangú szomszédossági követelmények teljesítésének előresorolása révén. Összetartozó szavak közé (pl. *a lány*) így kerülhetnek még jobban „odavágyódó" szavak (*a büszke magyar lány*), hozva esetleg függelékeiket is (*a (két (jóképű)) bátyjára büszke magyar lány*). A morfotaxist hasonló rangparaméteres szomszédossági követelményekre bízhatjuk (pl. *kutyát* > *kutyáét* > *kutyámét*), azzal a szignifikáns könnyebbséggel (ami a fenti regularitási állítással korrelál [20]), hogy egy morféma nem hozhat „függelékeket" magával.

A „totálisan lexikalista morfológia" imént vázolt eszméje a nyelvtípusok közötti különbségeket egy ún. *kopredikációs hálózat* absztrakt ábrázolási szintjén [8] képes irrelevánssá tenni, hiszen mindegy, hogy a *vár-hat-l-ak* ige morfémái vagy az *I may wait for you* angol szósor szavai keresik-e egymást. Ezen a reprezentáción keresztül számítógépes *fordítást* is szeretnénk majd végezni [7], egy nem túl távoli jövőben.

A morféma forrású szemantikai reprezentációt egy mexikói konferencián mutattuk be [6], a Springer-kötet adta kapcsolódó lehetőséget pedig a GASG nyelvtantípus matematikai definíciójának publikálására használtuk fel, mintegy szabadalmaztatva ezzel.

4 A jelenlegi mondatelemzőnk

Az alábbiakban idézünk néhány sort abból a hatvanhatból, amit programunk az 1. sorban látható szósorról mint egy `gramm("Mari...")` . *célfüggvény* tartalmáról közöl.

Kezdjük a végén: a 28. sorban *grammatikusnak* nyilvánítja a mondatot; amit nem tenne, ha nem tudta volna azonosítani a morfémáit (2. sor) megfelelő morfofonológiai környezetekben [5]. A szintaxisból két sort emeltünk ki a kötőszó kezelésével kapcsolatosan: a *Mari* főnév (`1,1`: az első szó első morfémája) a kötőszó függeléke (4.), és ez a kötőszó fogja képviselni az alanyt az igével való relációban (6.). A 9. és a 10. sor ugyanezt már a szemantikai egységek formától függetlenedett viszonyait ábrázoló *kopredikációs* reprezentáció nyelvén fejezi ki.

```
1. Mari és Juli kereshetik Pétert.
2. LEXIKAI EGYSÉGEK: ...
3. SZINTAXIS:
4. gr("noun","regent","conj",1,1,2,1)
5. ...
6. gr("noun","regent","subj",2,1,4,1)
7. KOPREDIKÁCIÓS VISZONYOK:
8. copr("and",2,1,"Mary",1,1,1,0,"arg")
9. copr("look-for",4,1,"and",2,1,1,1,"arg")
```

```
10. ...
11. SZEMANTIKA:
12. provref("fixpoint",[e(4,2,1)])
13. provref("old",[r(1,1,1)])
14. pred("Mary",1,[r(1,1,1)])
15. provref("new",[r(2,1,1)])
16. provref("<or=",[r(2,1,1),e(4,1,1)])
17. pred("element",2,[r(1,1,1),r(2,1,1)])
18. pred("element",2,[r(3,1,1),r(2,1,1)])
19. provref("old",[r(3,1,1)])
20. pred("Julie",3,[r(3,1,1)])
21. provref("new",[e(4,1,1)])
22. pred("look-for",4,[e(4,1,1),r(2,1,1),r(5,1,1)])
23. provref("new",[e(4,2,1)])
24. provref("<",[e(4,2,1),e(4,1,1)])
25. pred("may",4,[e(4,2,1),e(4,1,1)])
26. provref("old",[r(5,1,1)])
27. pred("Peter",5,[r(5,1,1)])
28. yes
```

A diskurzus-szemantikai reprezentációt teljes terjedelmében közöltük (12-27.), de a DRT-ben [17, 2] járatlanok számára csak a lényeget tudjuk közvetíteni. A 25. sor szerint *lehetséges* egy e4111 szituáció, ami nem más, mint hogy r211 *keresi* r511-et (22.), azaz az ismertnek tekintett (26.) *Pétert* (27.). Az r211 referens pedig egy r111 referensű *Mariból* (14.) és az r311-es *Juliból* (20.) álló kételemű csoport (17-18.).

A csoport alanyú igék személyragozását részben a Bánréti [12: 747. o.] által javasolt módon programoztuk le, pl. így egyeztetve: *Én és te/ti/Péter kereshetjük Marit.* Olyan *többes számú* személyragot használunk tehát, amelynek személyjegye a legmagasabb rangú koordinált személyével azonos (az imént az *én* első személye érvényesült).

Fogas kérdéseket vet fel a csoportos cselekvés interpretálása is. A jelenlegi programverzió hátterében az a szemlélet áll, hogy az egyes mondatok szemantikai ábrázolatában általában elegendő a fent illusztrált *csoportolvasat*; elegendő tehát a diskurzusba való beágyazás során később dönteni arról, hogy van-e okunk a csoport tagjainak egyedileg meghatározható szerepet tulajdonítani (pl. *disztributív* vagy *kölcsönös* viszonyt). Egyetlen esetben feltételeztünk markánsan disztributív olvasatot, egy eddig nem említett egyeztetési minta esetében: ahol egyes számban marad az ige, pl. *Péter és János keresi Marit* (és nem *keresik*, ami erősen azt sugallná, hogy együtt teszik). Vessük össze a fentivel ezt a disztributív interpretációt:

```
... pred("look-for",4,[e(4,1,1),r(1,1,1),r(5,1,1)])
    pred("look-for",4,[e(4,1,2),r(3,1,1),r(5,1,1)]) ...
```

Talán annyit e röpke szemléltetés is igazolt, hogy programunk a szokásosnál jóval szélesebb spektrumú mondatelemzést képes nyújtani egy korántsem triviális nyelvi fragmentumon. Fragmentumunk folyamatos bővítését, rendszerünk más nyelvekre való kiterjesztését és további „intelligens" nyelvtechnológiai alkalmazások kifejlesztését tervezzük; a megvalósíthatóságra a garanciát az elméletileg korrekt, gyakorlatilag pedig végsőkig egyszerűsített processzálás jelenti. Lassan kinőve a kísérleti szakaszból a Prolog helyett egy korszerű adatbáziskezelő és egy mintaillesztésben maximálisan hatékony nyelvre készülünk áttérni, és keressük a lehetőségét egy tudásbázist szimuláló korpusz beiktatásának.

Hivatkozások

1. Alberti, G.: GASG: Minimal Syntax, Maximal Lexicon and PROLOG, Paper read at ALLC/ACH '98, July 9. In Hunyadi, L. (ed.): ALLC/ACH '98. KLTE, Debrecen (1998) 81-83
2. Alberti, G.: Lifelong Discourse Representation Structures, Gothenburg Papers in Computational Linguistics 00-5 (2000) 13-20
3. Alberti, G., Balogh, K., Kleiber, J.: GeLexi Project: Prolog Implementation of a Totally Lexicalist Grammar. In de Jongh, Zeevat, Nilsenova (eds.): Proc. of the Third and Fourth Tbilisi Symp. on Language, Logic and Computation. ILLC, Amsterdam, and Univ. Tbilisi (2002)
4. Alberti, G., Balogh, K., Kleiber, J.,Viszket, A.: A totális lexikalizmus elve és a GASG nyelvtan-modell. In Maleczki, M. (ed.): A mai magyar nyelv leírásának újabb módszerei V. Szegedi Tudományegyetem (2002) 193-218
5. Alberti, G., Balogh, K., Kleiber, J.,Viszket, A.: Towards a totally lexicalist morphology. Talk at 6th International Conference on the Structure of Hungarian (ICSH6), Düsseldorf, Germany (2002). To appear in Kenesei, I., Piñón, Ch. (eds.): Approaches to Hungarian 9
6. Alberti, G., Balogh, K., Kleiber, J.,Viszket, A.: Total Lexicalism and GASGrammars: A Direct Way to Semantics. In Gelbukh, A. (ed.): Proceedings of CICLing2003 (Mexico City). LNCS N2588. Springer-Verlag, Berlin Heidelberg New York (2003) 37-48
7. Alberti, G., Balogh, K., Kleiber, J.,Viszket, A.: A fordítás totálisan lexikalista megközelítése. MANYE, Számítógépes nyelvészeti szekció, Győr (2003). Megj. előtt.
8. Alberti, G., Kleiber, J.: Extraction of Discourse-Semantic Information... In Cunningham, H., Paskaleva, E., Bontcheva, K., Angelova, G. (eds.): Information Extraction for Slavonic and Other Central and Eastern European Languages. Borovets, Bulgaria (2003) 63-69
9. Balogh, K., Kleiber, J.: Egy lexikalista nyelvtan morfoszintaxisának PROLOG-implementációja. OTDK- díjas pályamunka, JGYTF, Szeged (2001)
10. Balogh, K., Kleiber, J.: Computational Benefits of a Totally Lexicalist Grammar. In Matoušek, V., Mautner, P. (eds.): Text, Speech and Dialogue, Proceedings of TSD2003. Springer-Verlag, Berlin Heidelberg New York (2003) 114-119
11. Balogh, K., Kleiber, J.: A Morphology Driven Parser for Hungarian. Talk at the 5th Int. Tbilisi Symp. on Language, Logic and Computation. Org. by ILLC, Amsterdam, and U. Tbilisi (2003)
12. Bánréti, Z.: A mellérendelés. Kiefer, F. (szerk.) Strukturális magyar nyelvtan I. Mondattan. Akadémiai, Budapest (1992) 715-796
13. Borsley, R. D.: Modern Phrase Structure Grammar. Blackwell, Oxford Cambridge (1996)
14. Bresnan, J.: Lexical Functional Syntax. Blackwell, Oxford (2000)
15. Chomsky, N.: Syntactic Structures. The Hague, Mouton (1957)
16. Chomsky, N. (ed.): The Minimalist Program. MIT Press, Cambridge, Mass. (1995)
17. van Eijck, J., Kamp, H.: Representing discourse in context. In van Benthem, J., ter Meulen, A. (ed.): Handbook of Logic and Language. Elsevier, Amsterdam, The MIT Press, Cambridge, Mass. (1997)
18. Joshi, A. K.: Starting with Complex Primitives Pays Off. In Gelbukh, A. (ed.): Proceedings of CICLing2003 (Mexico City). LNCS N2588. Springer-Verlag (2003) 1-10
19. Karttunen, L.: Radical Lexicalism. Report No. CSLI 86-68, Stanford (1986)
20. Karttunen, L.: Computing with Realizational Morphology In Gelbukh, A. (ed.): Proceedings of CICLing2003 (Mexico City). LNCS N2588. Springer-Verlag (2003) 203-214
21. Partee, B., ter Meulen, G.B., Wall, R.P.: Mathematical Methods in Linguistics. Kluwer Academic Publ. (1990)
22. Prószéky, G.: Megértéstámogatás és gépi fordítás: nyelvtechnológia a XXI. század elején. VIII. Országos (Centenáriumi) Neumann Kongresszus (2003)
23. Schubert, K.: Metataxis (Contractive Dependency Syntax for Machine Translation). Foris, Dordrecht (1987)